SORT 49 (2) July-December 2025, 1-34 DOI:

Optimism correction of the area under the ROC curve, with missing data

Susana Rafaela Martins[1], María del Carmen Iglesias-Pérez[2], Jacobo de Uña-Alvarez[2]

Abstract

The area under the ROC curve (AUC) plays an inportant role in the study of the predictive capacity of regression models. It is well known that an inflated AUC may result when the same data are used for training and testing the model. In this paper optimism correction of the AUC in the presence of missing data is investigated. Complete case analysis, inverse probability weighting and multiple imputation are employed to address the issue of missing data. For each of these approaches, split-sample, K-fold cross-validation and leave-one-out cross-validation are employed to correct for the optimism of the AUC. The methods are compared through intensive Monte Carlo simulations in the particular setting of binary regression. Results suggest that all estimators are consistent with the exception of complete case analysis, which may be biased when missing is not completely at random. In general, a combined application of multiple imputation and leave-one-out cross-validation is recommended.

MSC: *62G05, 62G09, and 92B15.*

Keywords: *Cross-validation, logistic regression, missing values, multiple imputation, prediction.*

1. Introduction

Predictive models are widely used in different areas. They are used to evaluate credit risk, evaluate fraud hypotheses, analyse the performance of a machine or identify the individual who is at greatest risk of illness (Fawcett, 2006; Pepe, 2003; Wishart et al., 2012; Quintana et al., 2014; Garcia-Gutierrez et al., 2017). A critical step in the application of predictive models is the assessment of their predictive ability, that is, the performance when applied to new individuals. In this work, we focus on binary regression models with missing values and on the study of their discrimination capacity (Steyerberg et al., 2001, 2003).

[1] Escola Superior de Desporto e Lazer, Instituto Politécnico de Viana do Castelo, Portugal.

[2] Department of Statistics and OR, Universidade de Vigo, Spain.

Received: February 2025

Accepted: July 2025

The Area Under the ROC Curve (AUC) is a discrimination measure that evaluates the overall ability of a model to correctly classify observations into two distinct classes. It represents an accuracy index for a classifier. When its value is 1, it indicates perfect discrimination capacity. If this value is 0.5, the discrimination capacity is null (Fan, Upadhye and Worster, 2006; Pepe, 2003).

In general, when the same data are used for training and testing the model the AUC is overestimated. In other words the AUC estimate can be optimistic; this is so-called apparent AUC (APP AUC). Several methods to correct for this optimism have been proposed in the context of complete data. Such methods include Split-sample (SS), K-fold cross-validation (KF), Leave-one-out (LOO) cross-validation or Bootstrap (Iparragirre, Barrio and Rodriguez-Alvarez, 2019; Airola et al., 2011; Smith et al., 2014; Austin and Steyerberg, 2017). However, in practice, the data could have missing information. Missing data in the response variable or in the covariables may have an impact in the performance of prediction models. In particular, when the individuals with missing information are removed from the sample, the so-called Complete Case (CC) analysis, the estimators of the regression coefficients and of discrimination measures such as the AUC may be biased. There exist some estimators of the AUC proposed to adjust biases caused by missing values. Inverse Probability Weighting (IPW) is commonly used to correct the selection bias when the analysis is restricted to cases with complete information. Alternatively, the analysis could be improved by using other methods for missing data, such as Multiple Imputation (MI) (Li et al., 2021; Cho, Matthews and Harel, 2019).

To sum up, corrections for the optimism of the AUC together with methods to properly handle missing data are needed in practice. As far as we know, only few studies in the literature addressed these two problems. Wahl et al. (2016) focus on MI combined with Bootstrap and KF methods; however, they do not include LOO method, nor consider other methodologies for missing data such as IPW or CC. On the other hand, Mertens, Banzato and de Wreede (2020) investigate the problem of calibrating a prediction model in the presence of missing data using MI. Nevertheless, the problem of the estimation of the AUC is not considered in this latter paper.

The goal of this work is the correction of optimism in the estimation of AUC in the presence of missing values. We compare various methodologies for handling missing data - MI, CC, and IPW – combined with different optimism correction methods – SS, KF, and LOO. In this sense, this study provides novel contributions to the topic of correcting the optimism of the AUC with missing data. Specifically, the performance of IPW with optimism correction is investigated for the first time. Also, the benefits of LOO when correcting for the optimism of the AUC in the missing data setting are explored.

The rest of the paper is organized as follows. In Chapter 2 the usual estimators for the AUC, corrections for their optimism with complete data, and existing adaptations of the empirical AUC to the context of missing data are presented. In Chapter 3 the methodologies to correct the optimism of AUC are adapted to missing data. The methods are compared through intensive Monte Carlo simulations in the particular setting of logistic

regression in Chapter 4. In Chapter 5 we present two real data illustrations. Finally, the main conclusions or our study are reported in Chapter 6.

2. Methods for complete data

2.1. AUC for Binary Regression

AUC is the most used metric to evaluated the performance of classification models, representing the area under the ROC curve. The ROC curve is a graphical representation of the trade-off between the true positive rate and the false positive rate of a binary classifier as its discrimination threshold is varied. In the general binary regression model we denote by Y the response variable, X the corresponding q-dimensional vector of covariates, and $\{(X_i,Y_i), 1 \leq i \leq n\}$ a random sample of (X,Y).

For a given classifier $p(X)$, AUC is the probability that $p(X)$ takes a larger value for an individual randomly drawn from the diseased population compared to an individual sampled from the healthy population:

$$AUC = P(p(X_1) < p(X_2)|Y_1 = 0, Y_2 = 1),$$

where $Y_1 = 0$ and $Y_2 = 1$ represent the nondiseased and diseased population, respectively.

Alternatively, AUC can be expressed as the expected value of an indicator function:

$$AUC = E(I(p(X_1) < p(X_2))|Y_1 = 0, Y_2 = 1)$$

where the indicator function $I(.)$ takes the value 1 if its argument is true or 0 otherwise (Pepe, 2003).

Without loss of generality, consider the logistic regression model $Y_i \sim Bernoulli(p(X_i))$ where

$$p(X_i) = P(Y_i = 1|X_i) = \frac{\exp(\beta^{\intercal} X_i)}{1+\exp(\beta^{\intercal} X_i)} \tag{1}$$

and β is the unknown vector of regression coefficients, which can be estimated from the data leading to the feasible predictor $\hat{p}(X)$. Then, the AUC estimator is given by

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{i \in D_0} \sum_{j \in D_1} [I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5 I(\hat{p}(X_i) = \hat{p}(X_j))] \tag{2}$$

where $D_0 = \{i|Y_i = 0\}$ and $D_1 = \{i|Y_i = 1\}$ represent the nondiseased and diseased individuals in the sample, respectively, and where n_0 and n_1 are their corresponding cardinalities. The empirical AUC in (2) is indeed equal to the two-sample Mann-Whitney-Wilcoxon statistic.

2.2. Correction for the optimism of the AUC

The empirical AUC could be inflated if the same data set is used to fit and to test the model. The most commonly used methods for correcting the optimism of the AUC with

complete data are briefly reviewed below. See, for instance, Iparragirre et al. (2019) and references therein for further details.

2.2.1. Split-sample (SS) cross-validation

In SS the sample is randomly divided into two subsamples: *training (train)* sample and *test* sample. Subsequently, the regression coefficients are estimated from the training sample. Using these coefficients, the prediction of the response variable $\hat{p}(X_i)$ for the *test* sample is calculated according to (1). Taking into account these estimated probabilities and the Y_i values in the *test* sample, the AUC is calculated according to equation (2). SS cross-validation is frequently used with two samples with equal size.

2.2.2. K-fold (KF) cross-validation

In KF cross-validation, the sample is divided into K subsamples of approximately similar sizes. The sub-sample S_k, $1 \leq k \leq K$, is the $test_k$ sample and the set with all of the others sub-samples is the $train_k$ sample. The regression model is estimated with the $train_k$ sample and then it is used to compute $\hat{p}(X_i)$ for the X_i in the $test_k$ sample. So, the respective AUC is calculated. This procedure is repeated for all S_k, resulting in K AUCs. The corrected AUC is calculated using the mean of the K-AUCs. In general, $K = 10$ is the most commonly used in the literature.

2.2.3. Leave-one-out (LOO) cross-validation

In LOO one observation is omitted from the initial set and the regression model is fitted from the remaining observations. The estimated model is used to compute $\hat{p}(X_i)$ for the X_i that was left out. This procedure is repeated for all observations in the sample, and the AUC is calculated comparing the estimated probabilities to the corresponding Y_i.

2.3. AUC with missing values

In the missing data setting the empirical AUC in (2) is no longer available and, therefore, some adjustments are needed. Given a vector (X_i, Y_i) of covariables X_i and variable response Y_i, in line with the approaches of Molenberghs and Kenward (2007) and Chen, Wan and Zhou (2015), we consider the corresponding vector (Z_i, Z_i^{mis}), where Z_i is a d-dimensional vector with $0 < d \leq q$ that is observed for all i's, while Z_i^{mis} represents the variables that may or may not be available for some i's. Note that the response variable will be included in the vector Z_i^{mis} when its value is missing for some individuals. Let R_i be the indicator of missing values, meaning,

$$R_i = \begin{cases} 1, & if \;\; Z_i^{mis} \;\; is \;\; observed \\ 0, & if \;\; Z_i^{mis} \;\; is \;\; missing \end{cases} \tag{3}$$

The data are considered missing completely at random (MCAR) when the probability of missing values is independent of both Z_i and Z_i^{mis}. Consequently, under

MCAR assumption, the probability of missingness is given by $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i, Z_i^{mis}) = P(R_i = 0)$. On the other hand, missing at random (MAR) occurs when the probability of missing values depends on Z_i. Under MAR assumption, the probability of missingness is $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i)$. Another mechanism of missing values is missing not at random (MNAR). Under MNAR the probability of missingness depends on both Z_i and Z_i^{mis}. That is, the probability of missingness is $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i, Z_i^{mis})$. MNAR scenarios are difficult since the missing mechanism depends on variables which are not always available, so external information may be needed in order to proceed.

2.3.1. Complete Case (CC)

CC analysis simply proceeds by deleting from the sample the individuals that have missing information. Following the same idea presented in Li et al. (2021), the expression of the CC version of the AUC is

$$\widehat{AUC}_{cc} = \frac{\sum_{i\in D_0}\sum_{j\in D_1}[I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5I(\hat{p}(X_i) = \hat{p}(X_j))]R_iR_j}{\sum_{i\in D_0} R_i \sum_{j\in D_1} R_j} \tag{4}$$

The sums $\sum_{i\in D_0} R_i$ and $\sum_{j\in D_1} R_j$ are the number of observations without missing values that are nondiseased and diseased, respectively. The estimator (4) is consistent under MCAR. However, it may be inconsistent in MAR scenarios; see for instance Li et al. (2021).

2.3.2. Inverse Probability Weighting (IPW)

IPW uses the inverse of the estimated probability that an individual has complete information to weight each observation and thus to correct the potential selection bias. Let $W_i = 1/P(R_i = 1|X_i, Y_i)$ be the inverse probability of the observation to be complete. Aligned with the idea of Li et al. (2021), we use logistic regression to build a model for $P(R_i = 1|X_i, Y_i) = P(R_i = 1|Z_i)$ conditional on the fully observed variates (under MAR assumption), and then to obtain the weight estimates $\hat{W}_i$. Then, the AUC IPW estimator is

$$\widehat{AUC}_{ipw} = \frac{\sum_{i\in D_0}\sum_{j\in D_1}[I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5I(\hat{p}(X_i) = \hat{p}(X_j))]R_i\hat{W}_iR_j\hat{W}_j}{\sum_{i\in D_0} R_i\hat{W}_i \sum_{j\in D_1} R_j\hat{W}_j} \tag{5}$$

The sums $\sum_{i\in D_0} R_iW_i$ and $\sum_{j\in D_1} R_jW_j$ are the weighted observations without missing values that are nondiseased and diseased, respectively. When the MAR-logistic model for the weights W_i is correctly specified, the estimator (5) is consistent (See Section 2.5 in Li et al. (2021)).

2.3.3. Multiple Imputation (MI)

MI replaces missing data with imputed values, resulting in multiple 'completed' datasets. Several approaches exist for performing imputation, such as the multivariate normal model developed by Schafer (1997) or the full conditional specification (FCS) method proposed by Raghunathan et al. (2001); van Buuren (2007), also known as "chained equations". The FCS is based on distributions of fully observed variables and is one of the most used in multiple imputation to estimate the distribution of partially observed variables (Carpenter and Smuk, 2021). FCS is expected to be consistent when the involved chained equations are correctly specified. The good practical behaviour of MI has been widely studied in the literature; see for instance Zhu and Raghunathan (2015); Carpenter and Smuk (2021).

In the case of AUC, M imputations are performed and M datasets with no missing data are obtained. For each of these sets, the respective AUC is estimated, AUC_m, according to the equation (2). The AUC resulting from this methodology, $\widehat{AUC}_{mi}$, is estimated by the average of the M AUCs:

$$\widehat{AUC}_{mi} = \frac{1}{M}\sum_{m=1}^{M} AUC_m. \tag{6}$$

In this paper, we adopted the fully conditional specification approach and selected $M = 5$ imputations, following common practice and the recommendation of van Buuren (2018), who notes that increasing M beyond 5 is unlikely to alter the substantive conclusions. Since MI yields several completed datasets, we report the mean AUC across all imputed sets. Note that the objective in this study is to evaluate and compare the predictive performance of different methods for handling missing data and optimism correction, specifically, using the AUC as the parameter of interest. To estimate the AUC under MI, we adopt the common approach of computing it separately for each imputed dataset and then averaging the resulting AUCs. This strategy is consistent with recommendations in the literature (e.g., Wahl et al. (2016); Mertens et al. (2020)), and is preferred when the interest lies in assessing model performance rather than interpreting coefficients. It is worth noting that if one were interested in reporting a final model for implementation, the appropriate approach would be to pool the regression coefficients across imputations using Rubin's rules (Rubin, 1987). A single model could then be derived for interpretation purposes, and its AUC could be computed. However, this is not the focus of the current study.

3. Correction for the optimism of the AUC with missing values

Two general approaches to correct for the optimism of the AUC with missing data are possible. The first one corrects first for optimism and then deals with the missingness issue. This approach is feasible when using SS cross-validation or KF cross-validation, among other methods. However, the approach fails for LOO cross-validation, since prediction for an individual observation is not possible when some covariates are missing.

The second approach solves first the missingness issue and then proceeds to correct for optimism. This approach works for all the methods, and it will be employed in our research.

3.1. SS cross-validation for CC, IPW and MI methods

3.1.1. SS cross-validation for CC analysis

Taking into account the CC methodology, the set of complete observations is considered and this set is partitioned in half into *trainc* and *testc* subsamples. Using the set *trainc* the regression parameters are estimated and, from this estimated model, predictions are obtained for the *testc* set. Using the response values of *testc*, and their predictions, the AUC, which we call AUC_{cc-ss}, is calculated using (4).

3.1.2. SS cross-validation for IPW

In the IPW methodology, one estimates the weights, $W_i = 1/P(R_i = 1|Z_i)$, by plugging in a consistent estimator for the non-missing probability $P(R_i = 1|Z_i)$. Then, the dataset is divided into two sets with the same size, *trainc* and *testc*, and the respective pre-estimated weights are considered.With the *trainc* data set and the respective pre-calculated weights, a weighted logistic model is built and predictions for the *testc* sample are obtained. The corresponding AUC is calculated taking into account formula (5) and using the response values, pre-estimated weights and predicted values of the *testc* set. To be more specific, $X_i, X_j, W_i, W_j, R_i, R_j, D_0$ and D_1 are related to *testc*, while $\hat{p}$ is estimated using weighted logistic regression with *trainc* sample, and evaluated in the *testc* sample.

3.1.3. SS cross-validation for MI

In the case of MI, M (we take $M = 5$ in the simulations below) imputations are performed to construct M complete datasets. For each imputation, one splits the full set into two, *mictrain* and *mictest* subsets. Then, the regression coefficients are estimated from each mictrain set. Using the estimated regression model, the predicted values of the respective *mictest* sets are calculated and the corresponding AUCs are obtained from (2). The final AUC, AUC_{mi}, is defined as the mean of the M AUCs obtained.

3.2. KF cross-validation for CC, IPW and MI methods

3.2.1. KF cross-validation for CC analysis

Taking into account the CC approach, only the complete observations are considered. The set of complete observations is divided into K $(K = 10)$ sets and the sets $trainc_k$ and $testc_k$ are obtained as described in Section 2.2.2. Using the set $trainc_k$ the regression parameters are estimated. Based on this estimated model, predictions are obtained for

the $testc_k$ sample. With the true outcomes of each $testc_k$ and their predictions, each AUC is calculated, AUC_{cc_k}, according to (4). Finally the K, AUC_{cc_k} are averaged.

3.2.2. KF cross-validation for IPW

The data set is divided into K (we take $K = 10$ in the simulations below) sets, $train_k$ and $test_k$. The method proceeds first as described in Section 3.1.2 for each of the K folds of the dataset and then the final AUC is obtained by averaging.

3.2.3. KF cross-validation for MI

The method proceeds first as described in Section 3.1.3 for each of the K folds of the dataset and then the final AUC is obtained by averaging.

3.3. LOO cross-validation for CC, IPW and MI methods

3.3.1. LOO cross-validation for CC analysis

LOO cross-validation for CC proceeds just as described in 2.2.3 considering only the complete observations.

3.3.2. LOO cross-validation for IPW

In the IPW method, the weights are first estimated for all individuals in the sample. These weights correspond to the inverse of the estimated probability of having complete data, typically obtained by fitting a logistic model to the missingness indicator R_i, conditional on the fully observed variables (i.e., estimating $\hat{W}_i = 1/\hat{P}(R_i = 1|Z_i)$). After estimating the weights, leave-one-out cross-validation is applied. Each observation i, $1 \leq i \leq n$, is considered as the *test* set, while the remaining $n-1$ observations form the corresponding train set. Using the train set and the respective pre-estimated weights, a weighted logistic regression model is fitted to predict the outcome variable Y. The fitted model is then used to compute the predicted probability for the omitted observation i. This process is repeated for all n observations, resulting in a set of n predicted probabilities. Finally, the AUC is computed using these predictions and the observed outcomes, according to equation (5).

3.3.3. LOO cross-validation for MI

For the MI methodology, missing data are imputed and M complete samples are constructed. For each of this samples train and test sets are defined. Each observation i, $1 \leq i \leq n$ is considered the test sample, and the original set without this observation i is the train. The regression model is estimated from each of the train samples. Each estimated regession model is used to obtain the prediction of the outcome for the corresponding test sample. This results in M AUCs, one for each imputed dataset. The final AUC is obtained by averaging.

4. Simulation study

In this section the performance of optimism correction methods with missing values, as introduced in Section 3, is investigated through simulations. The setting is that of logistic regression. The goal is to identify the best methods to correct for missing values and for the optimism of the AUC. We compare the AUC estimated by each combination of methods to the "true" out-of-sample AUC associated to the fitted logistic regression models. As mentioned, the combination of methods involves a method to correct for data missingness (*mm*) and a method to correct for the optimism of the AUC (*om*). The "true" out-of-sample AUC represents the true discriminatory ability of the models, according to a particular missing method, when applied to new data without missingness. We consider various factors that might affect the methods' performance, including the sample size and disease prevalence as in Iparragirre et al. (2019). Inspired by Li et al. (2021), we considered different scenarios of missingness. Details are given in the next section.

4.1. Simulation design

In the simulation study two independent samples $\{X_i, Y_i\}_{i=1}^{n}$ and $\{X_l, Y_l\}_{l=1}^{N}$, *ndata* and *bigdata* say, from the population vector (X, Y) were generated, where Y was the binary response variable and X was a vector of eight covariates. The steps to simulate each (X_i, Y_i) were the following.

- Draw a Bernoulli (*prev*) variable, η_i,
- Given η_i, draw X_i from a multivariante Normal distribution with independent components with standard deviation 0.6 as follows:
  - If $\eta_i = 0$ the components of X_i were zero mean.
  - If $\eta_i = 1$ the vector mean of X_i was $(0.6, 0.55, 0.5, 0.45, 0.4, 0.3, 0.25, 0.2)$.
- Draw Y_i from a Bernoulli(π_i) distribution, where $\pi_i = \frac{\exp(\beta^T X_i)}{1+\exp(\beta^T X_i)}$, where β is the true vector of regression coefficients.

The prevalence values (*prev*) in the simulations were 0.1, 0.2 and 0.5. The true regression coefficient vector was
$\beta = [-2.5082, 0.5625, 0.4375, 0.3125, 0.1875, 0.0625, -0.1875, -0.3125, -0.4375]$.
Three different sizes were considered for *ndata*, $n = \{250, 500, 2000\}$. The size of *bigdata* was $N = 50000$ and $T = 500$ Monte Carlo trials were performed. Our simulation design resembles that in Iparragirre et al. (2019). A novelty here is the data missingness, which was introduced for a single covariate, namely X_1, according to different scenarios:

- S1: MCAR with missing probability $P(R = 0) = 0.5$.

- S2: MAR with missing probability depending on X_2 and X_3: $P(R=0|X_2,X_3) = 1/(1+exp(-0.5+2X_2-X_3)$
- S3: MAR with missing probability depending on X_2 and Y: $P(R=0|X_2,Y) = 1/(1+exp(-0.5+2X_2+1.5Y))$

On average, the probability of missingness in the MAR scenarios was approximately 0.5. To fit the logistic regression model we used the `glm` function of `R` with binomial family (R Core Team, 2013). The weights for IPW were also computed by fitting a logistic regression model with R function `glm`. For MI we used the function `mice` of the package in `R` with same name (van Buuren and Groothuis-Oudshoorn, 2011). We used $M = 5$ imputed datasets, applying the `norm` method for imputing variable X_1. To estimate the AUC, equations (4), (5) and (6) were implemented.

We evaluated the AUC on each Monte Carlo trial and then we computed the Monte Carlo average, bias and mean square error (MSE) as follows:

$$AUC_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}_{t;mm,om}) \tag{7}$$

$$Bias_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}^{n}_{t;mm,om} - \widehat{AUC}^{N}_{t;mm}) \tag{8}$$

$$MSE_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}^{n}_{t;mm,om} - \widehat{AUC}^{N}_{t;mm})^2 \tag{9}$$

where $\widehat{AUC}_{t;mm,om}$ denotes the generic estimator $\widehat{AUC}$ based on the different combination of methods, when computed from the t-th Monte Carlo trial, and where the upper index identify the respective sample. Note that missing method, $mm \in \{CC, IPW, MI\}$ is always the first one and the optimism method is the second, $om \in \{SS, KF, LOO\}$. See details in appendix A. The $\widehat{AUC}^{N}_{t;mm}$ is the out-of-sample AUC obtained with particular missing method *mm*; this is the target, and it varies from trial to trial since the fitted model varies too. We decided to generate a large sample (*bigdata*), because we are interested in an estimate with good accuracy and precision. This idea, used in Iparragirre et al. (2019), was previously considered by other authors (Austin and Steyerberg, 2017; Hsu and Chen, 2016; Smith et al., 2014; Steyerberg et al., 2001; Yan, Tian and Liu, 2015). It is important to mention that for *bigdata* no missing scenario was created, the sample was always complete.

4.2. Results

The simulation results are reported in Tables 1-4, and graphically summarized in Figures 2-4. From these results it is seen that the apparent AUC (denoted by APP in Tables 1-4 and Figures 2-4) overestimates the target. This was expected, since the apparent AUC

measures the discriminatory capacity of the model on the very data that were used to fit the model. The overestimation issue is much more evident with a small sample size; with $n = 2000$ the data become almost fully representative of the target population, so the issue vanishes to some extent.

Table 1 shows the AUC optimism corrections for the full data, which are not available in practice but are interesting for comparison purposes. The LOO method always presents the smallest MSE, sometimes equalled by KF. Regarding the bias, the closest to zero is always presented by KF, followed by LOO and finally SS, with the latter methods tending to underestimate the target (negative bias).

Table 1. *Optimism corrections for the AUC and respective bias and MSE. No missing data scenario.*

		$n = 250$			$n = 500$			$n = 2000$		
prev	method	AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7645	0.0924	0.0109	0.7351	0.0465	0.0034	0.7180	0.0130	0.0005
	SS	0.6454	-0.0267	0.0090	0.6697	-0.0189	0.0039	0.6988	-0.0063	0.0008
	KF	0.6711	-0.0010	0.0063	0.6863	-0.0023	0.0023	0.7045	-0.0006	0.0004
	LOO	0.6528	-0.0193	0.0057	0.6756	-0.0130	0.0022	0.7025	-0.0026	0.0004
0.2	APP	0.7427	0.0572	0.0047	0.7284	0.0307	0.0016	0.7149	0.0071	0.0003
	SS	0.6642	-0.0214	0.0051	0.6873	-0.0104	0.0020	0.7035	-0.0044	0.0005
	KF	0.6850	-0.0005	0.0028	0.6986	0.0009	0.0010	0.7070	-0.0009	0.0003
	LOO	0.6740	-0.0115	0.0026	0.6931	-0.0045	0.0010	0.7058	-0.0020	0.0003
0.5	APP	0.7317	0.0380	0.0024	0.7223	0.0203	0.0009	0.7147	0.0056	0.0001
	SS	0.6794	-0.0144	0.0028	0.6960	-0.0059	0.0012	0.7075	-0.0016	0.0002
	KF	0.6920	-0.0018	0.0015	0.7021	0.0001	0.0006	0.7095	0.0004	0.0001
	LOO	0.6853	-0.0085	0.0015	0.6985	-0.0034	0.0006	0.7087	-0.0005	0.0001

Table 2 corresponds to MCAR scenario S1. The CC and IPW methodologies provide very similar estimates to each other with a worse performance than MI, which reports the smallest MSE in all cases. In addition, the APP value with CC and IPW is greater than APP with MI (larger positive bias) and the CC/IPW corrections (SS, KF, LOO) overcorrect optimism (lower values) compared to MI corrections, especially with small prevalences and sample sizes. These results are illustrated in Figure 1, for $n = 500$ and $prev = 0.2$. In Figure 1, the three blocks of boxplots indicate the results of CC (left), IPW(middle) and MI (rigth) estimators, respectively, and the first boxplot of each block corresponds to the out-of-sample AUC (defined in Section 4) which is the target. Labels such as CC-OOS, CC-APP, CC-SS, ... indicate the specific estimation method used within each group. Considering MI, LOO is the correction method that presents the smallest MSE and the bias closest to zero. See Figure 2 for relative results on the bias of the several optimism correction methods when applied to MI with an increasing sample size. Note that from Table 2 to Table 4, the lowest MSE is shown in bold and the lowest bias is underlined.

Table 2. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.8058	0.1674	0.0330	0.8079	0.1708	0.0343	0.7689	0.1071	0.0142
	SS	0.6024	-0.0360	0.0209	0.6008	-0.0362	0.0215	0.6508	-0.0110	0.0083
	KF	0.6345	-0.0039	0.0155	0.6317	-0.0053	0.0163	0.6822	0.0204	0.0066
	LOO	0.5931	-0.0453	0.0171	0.5870	-0.0500	0.0188	0.6607	-0.0010	**0.0056**
0.2	APP	0.7688	0.1070	0.0140	0.7704	0.1100	0.0147	0.7469	0.0687	0.0063
	SS	0.6341	-0.0277	0.0090	0.6340	-0.0264	0.0088	0.6682	-0.0100	0.0048
	KF	0.6622	0.0004	0.0066	0.6609	0.0005	0.0069	0.6918	0.0136	0.0032
	LOO	0.6394	-0.0224	0.0063	0.6369	-0.0235	0.0066	0.6794	0.0012	**0.0029**
0.5	APP	0.7518	0.0749	0.0076	0.7528	0.0767	0.0080	0.7352	0.0474	0.0033
	SS	0.6533	-0.0235	0.0063	0.6518	-0.0243	0.0067	0.6831	-0.0047	0.0029
	KF	0.6747	-0.0022	0.0041	0.6724	-0.0037	0.0044	0.6960	0.0082	0.0016
	LOO	0.6622	-0.0147	0.0041	0.6605	-0.0156	0.0044	0.6893	0.0016	**0.0015**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7604	0.0916	0.0106	0.7612	0.0929	0.0109	0.7417	0.0580	0.0048
	SS	0.6368	-0.0320	0.0101	0.6364	-0.0319	0.0101	0.6771	-0.0066	0.0037
	KF	0.6655	-0.0033	0.0059	0.6648	-0.0035	0.0060	0.6914	0.0078	0.0027
	LOO	0.6470	-0.0218	0.0054	0.6457	-0.0225	0.0055	0.6831	-0.0006	**0.0024**
0.2	APP	0.7402	0.0555	0.0044	0.7405	0.0562	0.0045	0.7294	0.0353	0.0021
	SS	0.6626	-0.0221	0.0048	0.6617	-0.0227	0.0050	0.6875	-0.0066	0.0019
	KF	0.6835	-0.0012	0.0025	0.6836	-0.0008	0.0026	0.6991	0.0050	0.0012
	LOO	0.6722	-0.0125	0.0024	0.6713	-0.0130	0.0025	0.6943	0.0002	**0.0012**
0.5	APP	0.7325	0.0395	0.0025	0.7327	0.0398	0.0025	0.7241	0.0245	0.0011
	SS	0.6803	-0.0127	0.0029	0.6792	-0.0136	0.0031	0.6942	-0.0053	0.0012
	KF	0.6929	-0.0002	0.0015	0.6924	-0.0005	0.0015	0.7043	0.0047	0.0007
	LOO	0.6862	-0.0069	0.0014	0.6856	-0.0073	0.0015	0.7005	0.0010	**0.0006**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7228	0.0236	0.0012	0.7227	0.0235	0.0012	0.7207	0.0168	0.0007
	SS	0.6863	-0.0129	0.0019	0.6860	-0.0132	0.0019	0.7016	-0.0023	0.0008
	KF	0.6962	-0.0031	0.0009	0.6960	-0.0032	0.0009	0.7073	0.0034	**0.0005**
	LOO	0.6921	-0.0071	0.0009	0.6918	-0.0074	0.0009	0.7055	0.0015	**0.0005**
0.2	APP	0.7185	0.0140	0.0006	0.7184	0.0139	0.0006	0.7154	0.0083	**0.0003**
	SS	0.6970	-0.0075	0.0010	0.6969	-0.0076	0.0010	0.7043	-0.0028	0.0005
	KF	0.7025	-0.0020	0.0005	0.7023	-0.0021	0.0005	0.7076	0.0006	**0.0003**
	LOO	0.7004	-0.0041	0.0005	0.7003	-0.0042	0.0005	0.7063	-0.0007	**0.0003**
0.5	APP	0.7161	0.0094	0.0004	0.7161	0.0095	0.00044	0.7143	0.0060	0.0002
	SS	0.7010	-0.0057	0.0006	0.7010	-0.0057	0.0006	0.7071	-0.0012	0.0003
	KF	0.7059	-0.0008	0.0003	0.7058	-0.0008	0.0003	0.7090	0.0007	**0.0002**
	LOO	0.7041	-0.0026	0.0003	0.7040	-0.0026	0.0003	0.7083	0.0000	**0.0002**

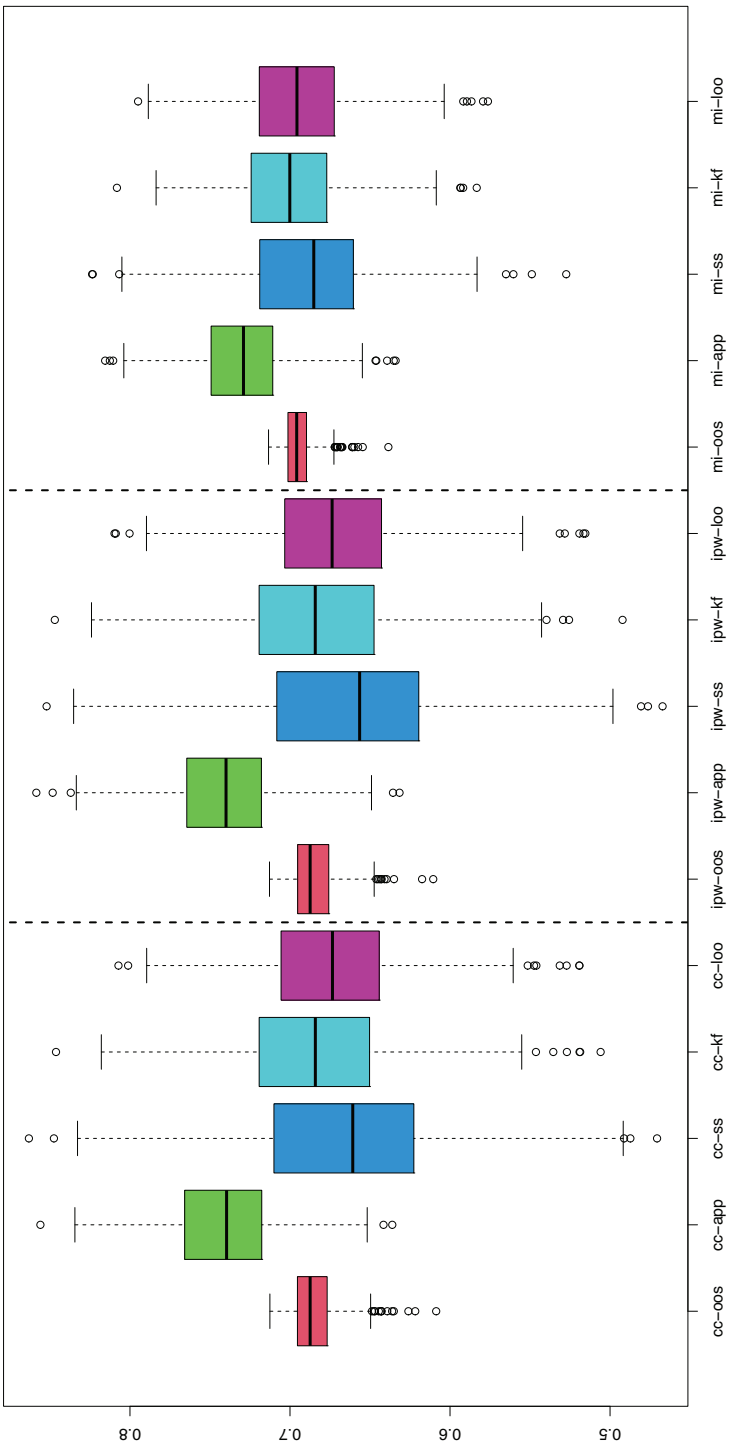

Figure 1. *Boxplot of AUC estimates with* $n = 500$ *and* $prev = 0.2$ *in the MCAR scenario S1 : CC left panel, IPW middle panel, and MI right panel.*

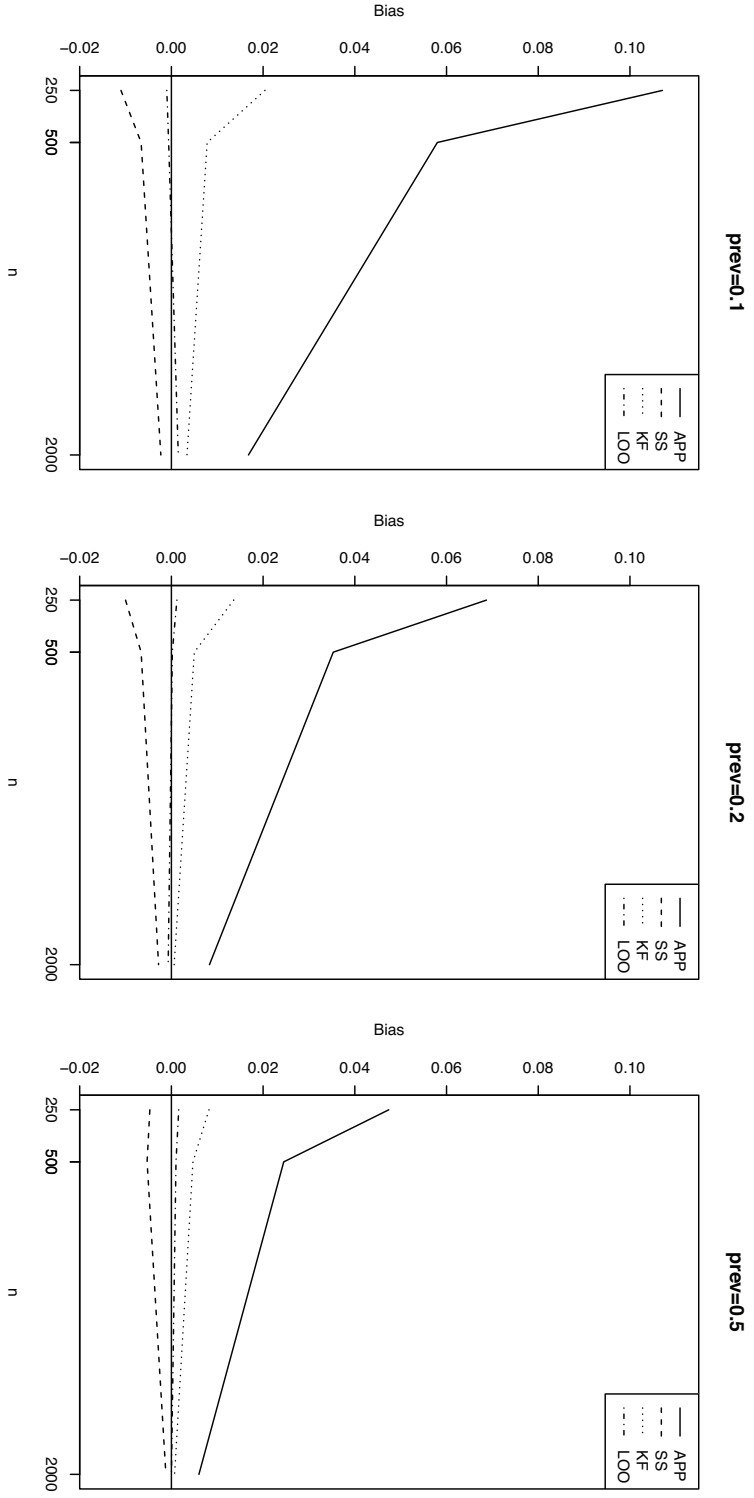

Figure 2. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MCAR scenario S1.*

Table 3. *Optimism corrections for the AUC and respective bias and MSE. MAR scenario S2.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7972	0.1582	0.0306	0.8455	0.2123	0.0506	0.7679	0.1084	0.0142
	SS	0.5855	-0.0534	0.0225	0.6026	-0.0306	0.0264	0.6590	-0.0005	0.0077
	KF	0.6221	-0.0168	0.0161	0.6487	0.0154	0.0149	0.6786	0.0191	0.0061
	LOO	0.5807	-0.0582	0.0181	0.5912	-0.0420	0.0218	0.6582	-0.0013	**0.0053**
0.2	APP	0.7602	0.1006	0.0130	0.8108	0.1648	0.0309	0.7437	0.0663	0.0057
	SS	0.6189	-0.0408	0.0113	0.6114	-0.0345	0.0179	0.6691	-0.0083	0.0041
	KF	0.6434	-0.0162	0.0077	0.6525	0.0066	0.0085	0.6857	0.0082	0.0026
	LOO	0.6199	-0.0398	0.0083	0.6085	-0.0375	0.0150	0.6750	-0.0024	**0.0024**
0.5	APP	0.7478	0.0748	0.0077	0.7867	0.1313	0.0207	0.7367	0.0504	0.0036
	SS	0.6395	-0.0335	0.0078	0.6215	-0.0339	0.0154	0.6872	0.0009	0.0028
	KF	0.6640	-0.0090	0.0050	0.6442	-0.0113	0.0071	0.6976	0.0112	0.0017
	LOO	0.6492	-0.0238	0.0048	0.6221	-0.0334	0.0123	0.6911	0.0047	**0.0016**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7554	0.0873	0.0101	0.7986	0.1441	0.0242	0.7415	0.0582	0.0046
	SS	0.6405	-0.0277	0.0090	0.6454	-0.0092	0.0129	0.6759	-0.0074	0.0034
	KF	0.6604	-0.0078	0.0054	0.6811	0.0265	0.0075	0.6926	0.0093	0.0024
	LOO	0.6400	-0.0282	0.0062	0.6364	-0.0182	0.0110	0.6826	-0.0006	**0.0022**
0.2	APP	0.7339	0.0525	0.0042	0.7745	0.1075	0.0142	0.7312	0.0367	0.0022
	SS	0.6511	-0.0303	0.0053	0.6542	-0.0128	0.0089	0.6907	-0.0039	0.0019
	KF	0.6743	-0.0071	0.0030	0.6804	0.0134	0.0050	0.7017	0.0072	**0.0012**
	LOO	0.6620	-0.0194	0.0030	0.6565	-0.0105	0.0075	0.6968	0.0023	**0.0012**
0.5	APP	0.7290	0.0377	0.0024	0.7575	0.0808	0.0083	0.7241	0.0255	0.0013
	SS	0.6698	-0.0215	0.0032	0.6619	-0.0147	0.0060	0.6945	-0.0041	0.0013
	KF	0.6867	-0.0046	0.0016	0.6755	-0.0011	0.0029	0.7038	0.0052	**0.0008**
	LOO	0.6778	-0.0135	0.0017	0.6639	-0.0128	0.0045	0.7005	0.0019	**0.0008**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7181	0.0191	0.0010	0.7443	0.0564	0.0045	0.7191	0.0155	0.0007
	SS	0.6808	-0.0182	0.0020	0.6780	-0.0099	0.0047	0.7001	-0.0034	0.0009
	KF	0.6903	-0.0087	0.0010	0.6985	0.0106	0.0021	0.7061	0.0025	**0.0005**
	LOO	0.6863	-0.0127	0.0011	0.6811	-0.0068	0.0030	0.7036	0.0001	**0.0005**
0.2	APP	0.7117	0.0082	0.0005	0.7313	0.0350	0.0021	0.7149	0.0084	0.0003
	SS	0.6883	-0.0152	0.0012	0.6833	-0.0129	0.0027	0.7035	-0.0030	0.0005
	KF	0.6958	-0.0077	0.0006	0.6999	0.0037	0.0011	0.7071	0.0006	**0.0003**
	LOO	0.6925	-0.0110	0.0006	0.6890	-0.0072	0.0015	0.7058	-0.0007	**0.0003**
0.5	APP	0.7111	0.0050	0.0003	0.7241	0.0246	0.0012	0.7139	0.0058	0.0002
	SS	0.6946	-0.0115	0.0008	0.6906	-0.0088	0.0016	0.7067	-0.0014	0.0003
	KF	0.6996	-0.0066	0.0004	0.6960	-0.0035	0.0007	0.7088	0.0007	**0.0002**
	LOO	0.6977	-0.0085	0.0004	0.6931	-0.0063	0.0010	0.7078	-0.0003	**0.0002**

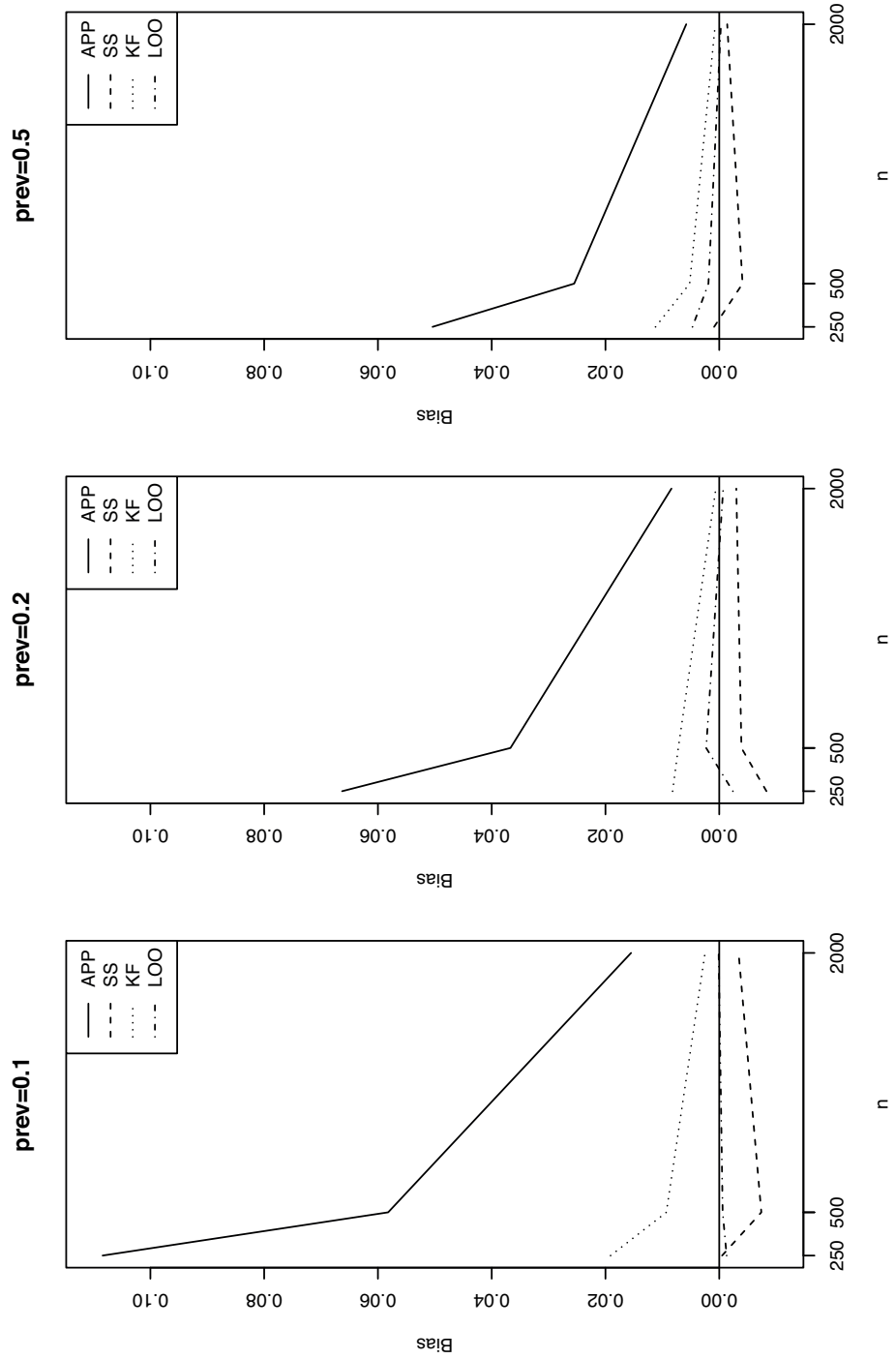

Figure 3. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MAR scenario S2.*

Table 4. *Optimism corrections for the AUC and respective bias and MSE. MAR scenario S3.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7453	0.1232	0.0208	0.8009	0.1562	0.0294	0.7683	0.1037	0.0131
	SS	0.5708	-0.0513	0.0176	0.6039	-0.0408	0.0213	0.6528	-0.0117	0.0082
	KF	0.5910	-0.0311	0.0147	0.6317	-0.0130	0.0124	0.6763	0.0117	0.0058
	LOO	0.5612	-0.0608	0.0160	0.6023	-0.0424	0.0167	0.6590	-0.0056	**0.0051**
0.2	APP	0.7193	0.0806	0.0095	0.7757	0.1147	0.0164	0.7436	0.0623	0.0055
	SS	0.5934	-0.0453	0.0099	0.6207	-0.0403	0.0148	0.6691	-0.0122	0.0046
	KF	0.6162	-0.0225	0.0067	0.6412	-0.0198	0.0069	0.6866	0.0053	0.0031
	LOO	0.5970	-0.0416	0.0075	0.6276	-0.0334	0.0099	0.6752	-0.0062	**0.0029**
0.5	APP	0.7055	0.0600	0.0061	0.7612	0.1014	0.0135	0.7320	0.0433	0.0029
	SS	0.6034	-0.0421	0.0079	0.6191	-0.0407	0.0131	0.6800	-0.0087	0.0024
	KF	0.6209	-0.0247	0.0054	0.6220	-0.0378	0.0068	0.6924	0.0037	**0.0016**
	LOO	0.6055	-0.0401	0.0062	0.6221	-0.0377	0.0106	0.6855	-0.0032	**0.0016**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7067	0.0578	0.0063	0.7646	0.0945	0.0116	0.7365	0.0515	0.0039
	SS	0.5997	-0.0491	0.0102	0.6431	-0.0270	0.0109	0.6702	-0.0148	0.0037
	KF	0.6190	-0.0298	0.0068	0.6647	-0.0054	0.0050	0.6861	0.0011	0.0027
	LOO	0.6022	-0.0466	0.0076	0.6472	-0.0229	0.0070	0.6763	-0.0087	**0.0022**
0.2	APP	0.6888	0.0285	0.0025	0.7484	0.0674	0.0064	0.7281	0.0332	0.0020
	SS	0.6137	-0.0466	0.0059	0.6534	-0.0275	0.0077	0.6889	-0.0061	0.0021
	KF	0.6287	-0.0316	0.0038	0.6661	-0.0149	0.0033	0.6980	0.0031	**0.0013**
	LOO	0.6196	-0.0407	0.0043	0.6628	-0.0181	0.0047	0.6928	-0.0021	**0.0013**
0.5	APP	0.6863	0.0200	0.0018	0.7380	0.0555	0.0048	0.7226	0.0230	0.0011
	SS	0.6254	-0.0409	0.0048	0.6518	-0.0307	0.0062	0.6948	-0.0049	0.0013
	KF	0.6407	-0.0256	0.0026	0.6549	-0.0276	0.0030	0.7024	0.0027	**0.0007**
	LOO	0.6329	-0.0334	0.0031	0.6611	-0.0215	0.0039	0.6989	-0.0008	**0.0007**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.6743	-0.0034	0.0008	0.7290	0.0320	0.0019	0.7186	0.0139	0.0006
	SS	0.6409	-0.0367	0.0030	0.6894	-0.0076	0.0026	0.6994	-0.0053	0.0008
	KF	0.6495	-0.0281	0.0017	0.6980	0.0010	0.0010	0.7051	0.0005	**0.0005**
	LOO	0.6450	-0.0326	0.0020	0.6925	-0.0044	0.0015	0.7032	-0.0014	**0.0004**
0.2	APP	0.6700	-0.0122	0.0006	0.7233	0.0213	0.0011	0.7163	0.0091	0.0003
	SS	0.6468	-0.0354	0.0022	0.6949	-0.0070	0.0018	0.7046	-0.0027	0.0004
	KF	0.6542	-0.0280	0.0013	0.6990	-0.0030	0.0007	0.7085	0.0013	**0.0002**
	LOO	0.6517	-0.0305	0.0015	0.6995	-0.0025	0.0009	0.7073	0.0001	**0.0002**
0.5	APP	0.6676	-0.0162	0.0006	0.7199	0.0179	0.0010	0.7143	0.0059	0.0002
	SS	0.6504	-0.0334	0.0018	0.6914	-0.0106	0.0018	0.7069	-0.0015	0.0003
	KF	0.6554	-0.0285	0.0012	0.6910	-0.0110	0.0008	0.7092	0.0008	**0.0002**
	LOO	0.6533	-0.0306	0.0013	0.6975	-0.0045	0.0010	0.7083	-0.0001	**0.0002**

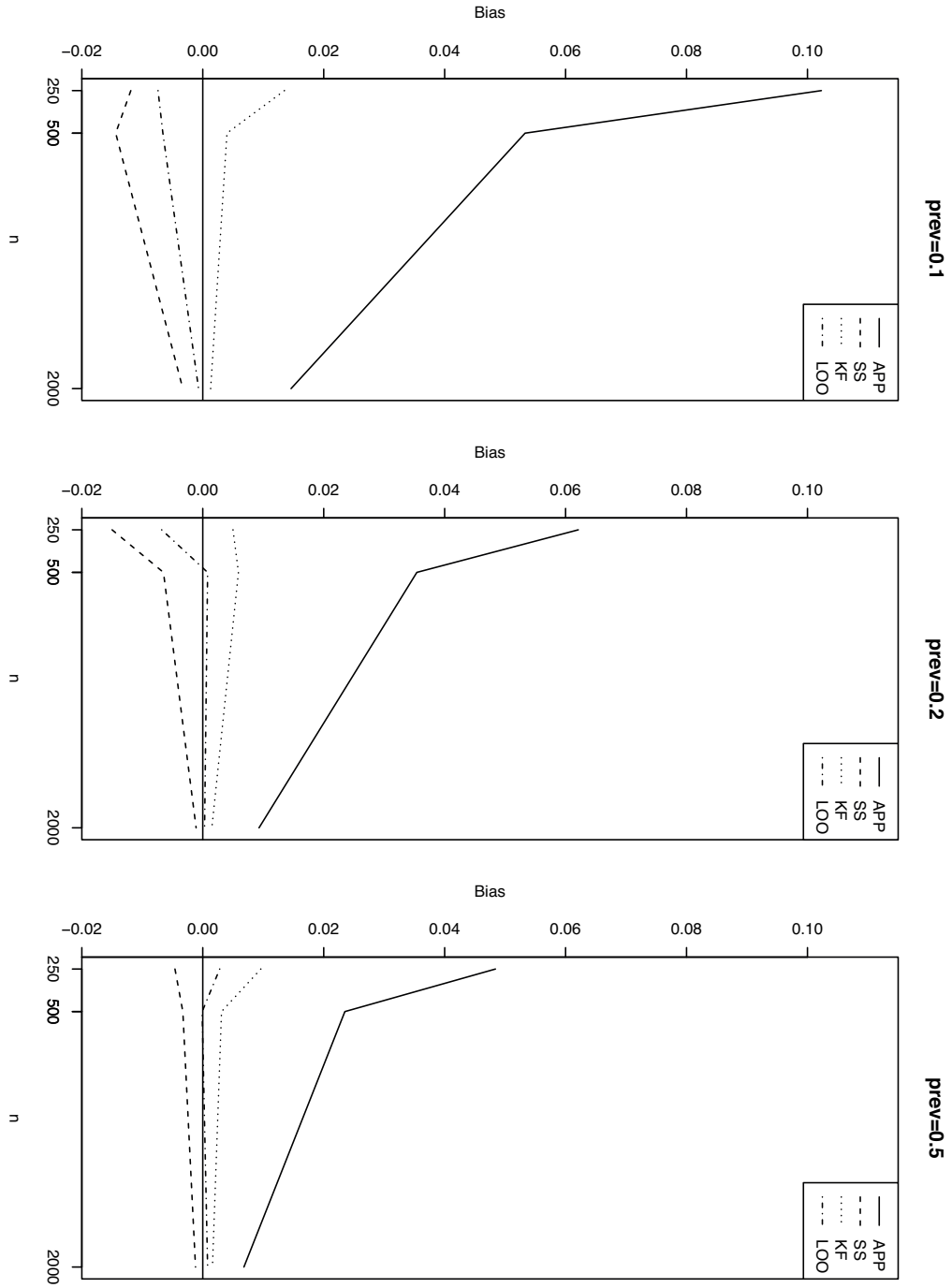

Figure 4. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MAR scenario S3.*

In the MAR scenario where the loss of the covariate X_1 depends only on other covariates (Table 3, scenario S2) the MI estimators have the smallest MSE's again. In this scenario the CC and IPW estimators provide different values, which are larger than MI for the APP AUC and lower than MI for the SS, KF and LOO corrections. All methods seem to be consistent, but IPW estimators have larger MSE's due to the noise in the estimation of the weights Wi in equation (5). In fact, some simulations (not shown here) presented quite high values of the apparent AUC due to denominators very close to zero. Considering MI, LOO is the correction method that presents the smallest MSE, equalled by KF as n increases. Moreover, the bias of LOO is between that of KF and SS corrections, with the latter method tending to overcorrect for the optimism (see Figure 3).

Table 4 shows the results of the MAR scenario S3 where the missingness of the covariate X_1 also depends on the response. It is notable that the CC methodology is not consistent. Both the values of the APP method and those of the corrections are below the values obtained with MI, IPW or the corresponding values for the full data (Table 1), maintaining these differences for $n = 2000$. Again, IPW presents a MSE larger than MI, with estimated values of the AUC larger than MI for the APP estimator and lower than MI for the SS, KF and LOO corrections, which get closer as n increases. Considering MI, LOO is the correction method that presents the smallest MSE, similar to KF as n increases. As in previous scenarios, the bias of LOO is between that of KF and SS (see Figure 4). But in this case, when the prevalence is low, the bias closest to zero is generally obtained by KF.

Following the suggestion of a referee, we performed additional simulations for the bootstrap method. The bootstrap method aims to correct for optimism in model performance by comparing how well a model performs on resampled data versus the original data. Specifically, we applied the bootstrap method in combination with multiple imputation, since multiple imputation was the method for missing data that exhibiting the best performance. First, missing values were imputed, generating m completed datasets. Then, for each imputed dataset j, $1 \leq j \leq m$, the following steps were carried out:

1. Fit a logistic regression model to the imputed dataset j and estimate the apparent AUC, denoted $\widehat{\mathrm{AUC}}_{\mathrm{app}}^{(j)}$.

2. For $b = 1, \ldots, B$ (we used $B = 500$ bootstrap iterations):

   (a) Draw a bootstrap sample from the imputed dataset j (with replacement).

   (b) Fit a logistic regression model to the bootstrap sample and compute its apparent AUC, $\widehat{\mathrm{AUC}}_{\mathrm{boot}}^{(j,b)}$.

   (c) Apply the fitted model to the full imputed dataset j, and compute the AUC, $\widehat{\mathrm{AUC}}_{\mathrm{imp}}^{(j,b)}$.

3. Estimate the optimism for imputation j as the average difference:

$$\widehat{O}^{(j)} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mathrm{AUC}}_{\mathrm{boot}}^{(j,b)} - \widehat{\mathrm{AUC}}_{\mathrm{imp}}^{(j,b)}\right)$$

4. Compute the optimism-corrected AUC for imputation j:

$$\widehat{\mathrm{AUC}}_{\mathrm{corr}}^{(j)} = \widehat{\mathrm{AUC}}_{\mathrm{app}}^{(j)} - \widehat{O}^{(j)}$$

Finally, the overall corrected AUC estimate was obtained by averaging the corrected AUCs across the $m = 5$ imputations:

$$\widehat{\mathrm{AUC}}_{\text{mi-boot}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\mathrm{AUC}}_{\mathrm{corr}}^{(j)}$$

The results of the bootstrap method are reported in Table 11, Appendix B. From Table 11 one may see that MI-Bootstrap yields the lowest MSE in scenarios S1 and S3, especially for small samples, although it has a substantial bias. In scenario S2, the bootstrap method shows a MSE similar to that of the other methods, but the bias tends to be larger. It is also worth noting that the bootstrap method is computationally demanding, particularly when combined with multiple imputation to handle missing data.

In the context of complete data, Iparragirre et al. (2019) found that the bootstrap method (and K-fold cross-validation with replication) were the methods with longer waiting times, althought the computational effort was generally affordable. Our experience is in fully agreement with that. For instance, for the chronic lymphocytic leukemia dataset analyzed in Section 5.2, the computational times ranged from approximately 25 or 21 seconds for the most time-consuming method (MI+Boot and MI+LOO respectively) to 0.02 seconds for the fastest method (IPW+SS). In particular, corrections based on MI multiplied by a factor between five and six the computational times attached to IPW. This should be taken into account when planning intensive data analyses or simulation studies.

4.3. Discussion

In general, for all simulated scenarios and with all methodologies for missing data, the apparent AUC overestimates the true AUC, so this estimator is optimistic. This optimism tends to disappear when the sample size increases. Also, all correction methods reduce optimism, with both the MSE and the bias of the corrected estimators approaching zero as the sample size increases. These results are in line with the studies for complete data by Steyerberg et al. (2001); Airola et al. (2011) as well as with the results for missing data in Wahl et al. (2016).

In the MCAR scenario, CC and IPW methods perform similarly and are consistent, although MI is more efficient. These results are in agreement with Li et al. (2021) regarding APP AUC estimation and, to our knowledge, are shown here for the first time for

AUC correction methods with missing data. Considering MI, the best AUC correction method in terms of bias and MSE is LOO.

In the MAR scenarios, MI and IPW estimators showed consistency, with MI being more efficient. However, CC estimators were clearly inconsistent in estimating the AUC specially when the loss mechanism depends on the response variable. These results are in agreement with Li et al. (2021), who compared CC, IPW and MI in estimating the apparent AUC. In this study, the same behaviour was observed in the correction methods to estimate the AUC. Considering MI, the LOO correction method presented the lowest MSE, sometimes tied with KF. In terms of bias, the bias of LOO is usually between that of KF and SS corrections, with SS and sometimes LOO tending to overcorrect (negative bias), although to a lesser extent than in the complete data scenario. The pessimistic behaviour of LOO and SS with complete data is in agreement with Austin and Steyerberg (2017) or Iparragirre et al. (2019). To our knowledge, this is the first time that LOO correction of AUC has been studied in the context of missing data. Although LOO has been found pessimistic in complete data, our simulations indicate that such pessimism may be attenuated in missing data scenarios.

In all simulated MCAR and MAR scenarios, MI estimators of AUC presented the smallest MSEs relative to their corresponding CC and IPW estimators. Our results agree with, and extend to the context of the AUC corrections, the results by Li et al. (2021), who highlight the better performance of MI over CC and IPW estimators of the AUC, and the limitations of the IPW method when multiple covariates are missing. Recent studies by Wahl et al. (2016) and Mertens et al. (2020), in the context of correcting optimism with missing data, seem to assume this premise because they focus only on the use of the MI methodology.

The imputations have been performed taking the response variable into account. In general, the literature on MI recommends including the outcome variable in the imputation models (Von Hippel (2007); Little (1992)). However, the impact of the imputation model in both the estimation of the AUC and the correction of its optimism has been less investigated. In our study, the absolute differences between the AUC estimates with MI (Tables 2-4) and the respective estimates with complete data (Table 1) are not relevant, being below or around 0.01 in all cases. The order in which the imputation and correction methods are combined is also important. For this, two general strategies are possible. One possibility is to perform the optimism correction first and then to apply MI. An alternative approach is to impute first and the to correct for the optimism (MI-OM strategy). According to Wahl et al. (2016) the estimates obtained by MI-OM are optimistically biased . In our study, we performed simulations based on MI-OM idea, but we did not observe a significant increase in the estimations when the response variable was included in the imputation. We opted for initiating the simulation process with MI (or other missing data methodologies) in order to apply LOO correction, since prediction for an individual observation is not possible when some covariates are missing.

As a complement, we conducted additional simulations under the MCAR mechanism with missing probabilities of 0.2 and 0.8 (Appendix B). These confirmed that the

optimism of the apparent AUC increases with higher missingness, and that the MI+LOO combination remains the most accurate and robust across all tested scenarios.

As a summary, we can say that, according to our simulation results, the MI methodology with the LOO correction method is the best combination because it has the smallest MSE and an ignorable bias.

5. Application to real data

5.1. Schoolchildren of Viana do Castelo dataset

The methods for correcting AUC optimism in the presence of missing data were applied to a case study on obesity in municipal schools in Viana do Castelo, based on the dataset by Rodrigues, Bezerra and Saraiva (2008). This dataset includes 229 children from northern Portugal. A logistic regression model was used to predict the International Obesity Task Force ($IOTF_{10}$) indicator based on physical examinations (ABD), past obesity status ($IOTF_4$), and sex. The binary response variable ($IOTF_{10}$) equals 1 for the presence of overweight or obesity and 0 for its absence. The disease prevalence in the dataset is approximately 0.18. The variable 'sex' does not have any missingness. Physical examinations had a 5% missingness rate, and both past and current obesity status had a 6% missingness rate. According Little test (Little (1992)) (p-value = 0.454), missing data occur under MCAR mechanism. The coefficients of prediction models and and respective p-values for CC and IPW methods are reported in Table 5. In these models all variables are significant considering a significance level of 0.1. For MI it is not possible to define a single prediction model because MI combines results from multiple imputed datasets, yielding a model that is a summary of models.

Table 5. *Coefficients and p-values of CC and IPW prediction models.*

	CC		IPW	
variable	coeficient	p-value	coeficient	p-value
intercept	- 0.0825	0.9348	-0.0819	0.9333
sex	- 0.8759	0.0661	-0.8759	0.0579
$IOTF_4$	2.9606	5.91e-10	2.9605	1.63e-10
ABD	- 0.0861	0.0047	-0.0861	0.0035

To estimate the weights in the IPW methodology, we considered the variable 'sex', which is the only complete variable, and the following logistic model was obtained:

$$logit(P(R=1)) = 2.70805 + 0.05407 \times sex$$

and variable sex was not significant (p-value = 0.923), that is according to the results of Little test.

Table 6. *AUC values in obesity case study according different methods.*

method	CC	IPW	MI
APP	0.8977	0.8977	0.8998
SS	0.8819	0.8818	0.8920
KF	0.8701	0.8701	0.8846
LOO	0.8759	0.8759	0.8976

In the case of Viana do Castelo study, Table 6, CC and IPW present similar results for the apparent AUC and its several corrections, while MI yields slightly higher AUC values compared to the others. KF is the method that produces the lowest AUC values, particularly for CC and IPW. In this data set, the SS results do not correspond to what one could expect given the results of our simulation study, since this method presents a higher AUC than other correction methods. Actually, Viana do Castelo study differs from the simulation scenarios in Section 4 in that the optimism of the AUC is almost ignorable. Still, one can compare these results with the MCAR scenario with $prev = 0.2$ and $n = 250$, as it is the most comparable scenario. In this dataset, the missing data mechanism is MCAR, the data size is $n = 229$, and the prevalence is $prev = 0.18$.

5.2. Chronic lymphocytic leukemia dataset

In this section we consider the chronic lymphocytic leukemia dataset provided by the European Society for Blood and Marrow Transplantation, previously analyzed by Schetelig et al. (2017). This dataset includes 694 patients and the same variables used by Mertens et al. (2020) in their study on the Brier score. A logistic regression model was employed to predict each individual's disease status (Status) based on patient-related variables. The predictors include age, performance status at transplantation (perfstat), cytogenetic abnormalities (cyto), remission status (remstat), prior treatments (asct), donor characteristics (donor), sex match (sexm) (between donor and patient), and clinical conditions (cond). The variable perfstat has four levels: Karnofsky 100, Karnofsky 90, Karnofsky 80, and Karnofsky $<= 70$. The variable remstat has three levels: CR, PR, and SD/PD. The variable cyto has four levels: del17p, del11q, other, and no abnormality. The variable asct is dichotomous: no prior ASCT and prior ASCT. The variable donor has three levels: matched related, matched UD, and partially mismatched UD. The variable sexm has four levels: PATmaleDONmale, PATmaleDONfemale, PATfemaleDONmale, and PATfemaleDONfemale. Finally, the variable cond has three levels: NMA, RIC, and MAC.

The binary response variable, Status, takes the value 1 for diseased individuals and 0 for healthy individuals. The prevalence of disease is around 0.27. The response variable (disease status) and the covariates age, prior treatments and donor characteristics are complete (no missing data). On other hand, performance status has 9% of missingness, remission status has 6%, cytogenetic abnormalities has 25%, and there is a 1%

missingness each for sex match and clinical conditions. According to Little test (p-value = 5.59e-6), we reject the null hypothesis of missing completely at random. Violation of MCAR assumption brings concerns on the results provided on the application of the CC approach.

The coefficients of prediction models and respective p-values of CC and IPW are those in Table 7. For CC, only perfstat and PR are significant, considering a significance level of 0.1. For IPW, additionally to these variables, age, matched UD and PATmale-DONfemale are significant, considering a significance level of 0.1. The results provided by CC are reasonable, despite of its potential inconsistency (MCAR assuption was rejected). Since the AUC estimated from MI is averaged along five different models, these are not reported in Table 7.

Table 7. *Coefficients and p-values of CC and IPW prediction models.*

	CC		IPW	
variable	coeficient	p-value	coeficient	p-value
Intercept	-1.6133	0.0004	-1.6098	9.98e-06
age	0.2002	0.1864	0.2216	0.0634
Karnofsky 90	0.2221	0.4786	0.2148	0.3843
Karnofsky 80	1.1209	0.0017	1.1627	3.84e-05
Karnofsky $<= 70$	2.0004	0.0023	1.9881	0.0002
PR	-0.6992	0.0522	-0.6782	0.0178
SD/PD	0.0805	0.8301	0.1477	0.6170
del11q	-0.1201	0.7050	-0.0578	0.8198
other	-0.1860	0.5382	-0.1522	0.5287
no abnormality	0.2029	0.6013	0.1997	0.5223
prior ASCT	-0.0879	0.8461	-0.0086	0.9773
matched UD	0.3866	0.1461	0.3891	0.0625
partially mismatched UD	0.4276	0.2513	0.4399	0.1434
PATmaleDONfemale	0.4477	0.1297	0.4665	0.0466
PATfemaleDONmale	-0.2596	0.4538	-0.1730	0.5290
PATfemaleDONfemale	-0.2298	0.5589	-0.2613	0.4101
RIC	0.3063	0.2719	0.3050	0.1668
MAC	-0.0461	0.9065	0.0029	0.9924

To estimate the weights in the IPW methodology, we only use the complete variables. The final logistic models is the following:

$$logit(P(R=1)) = 0.6385 + 0.2535 \times age - 0.9365 \times asct + 0.3722 \times \text{donor\_matched UD} + 0.3118 \times \text{donor\_partially mismatched UD} - 0.3082 \times \text{Status}$$

In the logistic model of weighting, all variables are significant considering a significance level of 0.1, including the outcome (disease status).

Table 8. *AUC values in leukemia case study according to different methods.*

method	CC	IPW	MI
APP	0.7009	0.7080	0.6860
SS	0.5818	0.5772	0.6156
KF	0.6185	0.6350	0.6342
LOO	0.6201	0.6281	0.6318

The estimated AUCs are reported in Table 8. In this data set, the optimism of the apparent AUC is evident, as it consistently presents the highest values across all missing data methodologies. CC and IPW give higher values than MI in the APP and lower in the corrections. This is in agreement with the simulation results, see Table 4. Also, CC (at least KF and LOO) seems to move further away from MI and IPW. We compare these results with the MAR scenario S3 with $prev = 0.2$ and $n = 500$ because it is the most similar scenario, due to the similar sample size and prevalence, and the fact that the missing probability depends on the outcome. Among the correction methods, SS yields the lowest AUC values but tends to underestimate the true AUC, according to simulations. In this data set, for all missing data approaches, KF and LOO methods produce close results.

The proportion of missing data is a crucial factor influencing the performance of both missing data handling methods and optimism correction techniques. In our simulation study, we evaluated scenarios with varying levels of missingness and observed that higher missing data rates accentuate the differences between methods. Specifically, approaches such as CC and IPW tend to lose reliability as missingness increases, while MI combined with LOO consistently exhibits strong performance across all settings. The real data applications corroborate these findings. In the Viana do Castelo dataset, where the overall missingness was relatively low (under 10%), all methods produced similar AUC estimates, and the impact of optimism was negligible. In contrast, the leukemia dataset, which presented higher missingness levels (up to 25% in some variables), revealed more pronounced differences between methods. In particular, CC corrections often underestimated the AUC and showed less stable results compared to MI or IPW. These findings are consistent with our simulation results and emphasize the need to consider the proportion of missing data when choosing an appropriate estimation strategy.

6. Main conclusions

In this paper, corrections for the optimism of the AUC in the presence of missing data have been investigated. All methods successfully corrected the optimism in the AUC. An exception was CC which, as expected, failed to provide unbiased estimations when the missingness is not completely at random. Among the several methods being compared, LOO achieved the lowest MSE. This is an interesting finding, since LOO has been previously reported as too pessimistic with complete data. Correction methods performed

particularly well with MI. In practice, we recommend using the combination of MI with LOO because of its good relative performance. Importantly, LOO method for optimism correction of the AUC with missing data had not been considered in the related literature.

The two real data illustrations provided in this manuscript cover two different situations that may appear in practice. For the Viana do Castelo study, the optimism of the AUC is negligible, and all methods roughly report the same result. Missingness can be assumed to be completely at random in this case. However, in the leukemia study the optimism of the AUC is evident, and choosing one or another method matters. Specifically, it has been seen how CC introduces some bias in this case, probably related to the fact that the missing probability depends on the outcome. For the leukemia data, IPW or MI methods, with KF or LOO optimism corrections, provided close results.

In addition to the main results reported for a missing value probability of 0.5, further simulations were conducted with lower (0.2) and higher (0.8) missing proportions, as presented in Appendix B. These results confirmed that the benefit of applying optimism correction methods becomes more evident as missingness increases. The MI+LOO combination consistently provided the lowest bias and MSE, even under high missingness levels, reinforcing its robustness and practical relevance.

Key Conclusions and Recommendations

- The MI method proved to be the most effective approach for handling missing data, because it consistently yielded the lowest bias and MSE across nearly all simulated scenarios.
- To correct optimism in AUC estimates in the presence of missing data, we recommend using MI combined with LOO cross validation.
- The KF cross validation method also produced competitive results and may serve as a viable alternative, particularly when computational efficiency is a concern.
- In MCAR settings with large sample sizes and high prevalence, CC combined with LOO or KF can be considered a reasonable alternative to MI.
- However, CC is not recommended in MAR scenarios, as it tends to introduce substantial bias.

Acknowledgments

The authors thank one anonymous reviewer and an Executive Editor for comments and suggestions which have served to improve the paper.

A. Details of methods

In this section we present a structured details of implemented methods to correct the AUC optimism with missing data.

- **Complete case analyses (CC)**
  - Apparent in complete case analyses (CC-APP)
    - Only the complete observations are considered. There are the observations with no missing values, that is, the observations with $R_i = 1$.
    - Fit the logistic model to Y depending on covariables.
    - Predicted values of this model are computed.
    - The AUC_{cc-app} is calculated, according to equation (4).
  - Split-sample validation in complete case analyses (CC-SS)
    - Considered observations with no missing values.
    - The sample with complete cases is divided into two: *trainc* and *testc*.
    - Using *trainc* the logistic model is fit to Y depending on covariables.
    - The predicted values of this model are computed to *testc*.
    - The AUC_{cc-ss} is calculated, according to equation (4).
  - K-fold cross-validation in complete case analyses (CC-KF)
    - Only considering the complete observations, that is the observations with $R_i = 1$.
    - Divide the sample into K sets.
    - For $k = 1$ define the subset $test_k$ and the $train_k$ that is the sample without the $test_k$ set.
    - Fit a logistic model to Y depending by the $train_k$.
    - Computed the predicted values of this model to $testc_k$.
    - The AUC_k is calculated, according to equation (4).
    - Repeat this for each $k \in 2,3,...,K$.
    - The AUC_{cc-kf} is the mean of the AUC_k.
  - Leave-one-out cross-validation in complete case analyses (CC-LOO)
    - Considering the set with no missing values, that is, the observations with $R_i = 1$.
    - Define $trainc_j$ and $testc_j$.
    - Using $train_j$ fit a logistic model to Y.
    - The predicted values of this model are computed to $testc_j$.

    - The set of predictions, was construct with all of predictions of $test_j$
    - The AUC_{cc-loo} is calculated, according to equation (4).

- **Inverse probability weighting (IPW)**
  - Apparent in inverse probability weighting (IPW-APP)
    - Define the weighting of each no missing observation by $1/p_i$, where $p_i = E(R_i|Y,X) = E(R_i|Z)$.
    - The observations with no missing values, the observations with $R_i = 1$ are considered.
    - Fit a logistic model to Y depending on covariables considering all complete observations and the respective weight.
    - The predicted values of this data set are computed.
    - The $AUC_{ipw-app}$ is calculated, according to equation (5).
  - Split-sample validation in inverse probability weighting (IPW-SS)
    - Define the weighting of each no missing observation similar to IPW-APP.
    - The observations with no missing values, the observations with $R_i = 1$, are considered and divided into two sets: *trainc* and *testc*.
    - Fit a logistic model to Y depending on covariables considering the observations of *trainc* set and the respective weight.
    - The predicted values of *testc* with this model are computed.
    - The AUC_{ipw-ss} is calculated, according to equation (5).
  - K-fold cross-validation in inverse probability weighting (IPW-KF)
    - Define the weighting of each no missing observation.
    - Considering only the observations with no missing values and divide the sample into K sets.
    - For $k = 1$ define the subset $test_k$ and the $train_k$ that is the sample without the $test_k$ set.
    - Fit a logistic model to Y considering the observations of $trainc_k$ set and their weights.
    - The predicted values of $testc_k$ with this model and the respective AUC, AUC_k, are computed.
    - Repeat this for each $k \in 2,3,...,K$.
    - The AUC_{ipw-kf} is the mean of the AUC_k.
  - Leave-one-out cross-validation in inverse probability weighting (IPW-LOO)
    - Estimate the weighting model.

  - Define $train_j$ and $test_j$.
  - Fit a logistic model to Y considering the $train_j$ set and the respectively weights.
  - Computed the predict value of $test_j$.
  - Repeat this for all values of j.
  - The $AUC_{ipw-loo}$ is computed to the set of predictions.

- **Multiple imputation (MI)**
  - Apparent in multiple imputation (MI-APP)
    - The original data was imputed m times, getting m full sets $micdata_i$, $i \in 1,2,..,m$.
    - For each one, the logistic model to Y depending on covariables was considered.
    - The predicted values to each set are computed.
    - The AUC of each m sets is calculated.
    - The AUC_{mi-app} is calculated, averaging the previous AUC.
  - Split-sample validation in multiple imputation (MI-SS)
    - The original data was imputed m times, getting m full sets $micdata_i$, $i \in 1,2,..,m$.
    - Each of this sets was is divided into two: $mictrain_i$ and $mictest_i$.
    - For each one, the logistic model to Y depending on covariables of $mictrain_i$ set was considered.
    - The predicted values to each set $mictest_i$ are computed.
    - The AUC of each *m test* sets is calculated.
    - The AUC_{mi-ss} is calculated, averaging the previous AUC.
  - K-fold cross-validation in multiple imputation (MI-KF)
    - The data set was imputed m times, getting m full sets $cdat_j$, $j \in 1,2,..,m$.
    - Divide each full set $cdat_j$ into K sets.
    - For $k = 1$ define the subset $mictest_{j_k}$ and the $mictrain_{j_k}$ that is the j complete set without the $mictest_{j_k}$.
    - The logistic model to Y of $mictrain_{j_k}$ set was considered.
    - The predicted values to each set $mictest_{j_k}$ and respective AUC, AUC_{j_k}, are computed.
    - The AUC_k is calculated averaging the AUC_{j_k}.
    - Repeat this for all values $k \in 2,3,...,K$.
    - The AUC_{mi-kf} is the mean of the AUC_k.

- Leave-one-out cross-validation in multiple imputation (MI-LOO)
  - Imputed the data set by m multiple imputations, getting m full sets.
  - To each set, define $mictrain_j$ and $mictest_j$.
  - Using $mictrain_j$ fit a logistic model to Y.
  - The predicted values of this model are computed to $mictest_j$.
  - The set of predictions, was construct with all of predictions of $mictest_j$ and the respective AUC is computed.
  - The AUC_{mi-loo} is the mean of the last m AUC's.

B. Additional simulation results

Table 9. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1 with a probability of missing 0.2.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7702	0.1099	0.0152	0.7708	0.1107	0.0154	0.7596	0.0922	0.0108
	SS	0.6283	-0.0320	0.0127	0.6277	-0.0325	0.0128	0.6407	-0.0266	0.0093
	KF	0.6563	-0.0040	0.0077	0.6546	-0.0055	0.0078	0.6671	-0.0003	0.0062
	LOO	0.6307	-0.0296	0.0082	0.6298	-0.0304	0.0083	0.6457	-0.0216	**0.0057**
0.2	APP	0.7469	0.0696	0.0068	0.7470	0.0698	0.0068	0.7406	0.0575	0.0048
	SS	0.6543	-0.0230	0.0068	0.6539	-0.0232	0.0069	0.6639	-0.0192	0.0047
	KF	0.6730	-0.0043	0.0042	0.6719	-0.0053	0.0043	0.6842	0.0011	0.0029
	LOO	0.6610	-0.0163	0.0039	0.6602	-0.0170	0.0039	0.6718	-0.0113	**0.0028**
0.5	APP	0.7382	0.0503	0.0037	0.7383	0.0505	0.0037	0.7324	0.0403	0.0026
	SS	0.6750	-0.0128	0.0032	0.6746	-0.0132	0.0032	0.6784	-0.0137	0.0027
	KF	0.6896	0.0017	0.0020	0.6891	0.0013	0.0020	0.6927	0.0006	0.0016
	LOO	0.6809	-0.0069	0.0019	0.6802	-0.0075	0.0019	0.6860	-0.0061	**0.0015**

Table 10. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1 with probability of missing 0.8.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.9012	0.3212	0.1153	0.9082	0.3287	0.1197	0.7936	0.1594	0.0300
	SS	0.5472	-0.0328	0.0488	0.5524	-0.0271	0.0500	0.6866	0.0525	0.0125
	KF	0.5624	-0.0176	0.0364	0.5614	-0.0180	0.0344	0.7135	0.0794	0.0142
	LOO	0.4666	-0.1134	0.0625	0.4655	-0.1140	0.0640	0.6960	0.0619	**0.0114**
0.2	APP	0.8457	0.2261	0.0587	0.8547	0.2382	0.0644	0.7640	0.1049	0.0140
	SS	0.5673	-0.0523	0.0281	0.5642	-0.0523	0.0294	0.6940	0.0349	0.0074
	KF	0.6136	-0.0060	0.0210	0.6120	-0.0045	0.0226	0.7109	0.0518	0.0074
	LOO	0.5456	-0.0740	0.0248	0.5372	-0.0793	0.0284	0.7012	0.0421	**0.0063**
0.5	APP	0.8018	0.1626	0.0308	0.8130	0.1785	0.0363	0.7499	0.0772	0.0079
	SS	0.5900	-0.0492	0.0187	0.5865	-0.0480	0.0197	0.6992	0.0265	0.0044
	KF	0.6288	-0.0104	0.0138	0.6247	-0.0098	0.0155	0.7130	0.0403	0.0042
	LOO	0.5972	-0.0420	0.0137	0.5933	-0.0412	0.0145	0.7064	0.0337	**0.0038**

Table 11. *Optimism corrections for the AUC and corresponding bias and MSE using MI and bootstrap across three missingness scenarios with approximately 0.5 probability of missing.*

	n	250			500			2000		
S	*prev*	AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
S1	0.1	0.6960	0.0365	0.0050	0.7009	0.0171	0.0021	0.7085	0.0044	0.0006
	0.2	0.6951	0.0174	0.0027	0.7049	0.0109	0.0011	0.7082	0.0014	0.0003
	0.5	0.7006	0.0137	0.0016	0.7042	0.0054	0.0007	0.7091	0.0011	0.0002
S2	0.1	0.6969	0.0404	0.0056	0.6990	0.0168	0.0021	0.7071	0.0035	0.0005
	0.2	0.6995	0.0229	0.0027	0.7014	0.0089	0.0013	0.7072	0.0002	0.0003
	0.5	0.7014	0.0149	0.0017	0.7055	0.0066	0.0008	0.7096	0.0016	0.0002
S3	0.1	0.6911	0.0263	0.0044	0.6952	0.0103	0.0018	0.7055	0.0012	0.0000
	0.2	0.6971	0.0164	0.0022	0.6997	0.0043	0.0009	0.70955	0.0026	0.0000
	0.5	0.6990	0.0109	0.0016	0.7055	0.0056	0.0008	0.7091	0.0009	0.0000

Funding

Work supported by the grants PID2020-118101GB-I00, Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033) and PID2023-148811NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Airola, A., T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis 55*(4), 1828–1844.

Austin, P. C. and E. W. Steyerberg (2017). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research 26*(2).

Carpenter, J. R. and M. Smuk (2021). Missing data: A statistical framework for practice. *Biometrical Journal 63*(5), 915–947.

Chen, X., A. T. K. Wan, and Y. Zhou (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association 110*(510), 723–741.

Cho, H., G. J. Matthews, and O. Harel (2019). Confidence intervals for the area under the receiver operating characteristic curve in the presence of ignorable missing data. *International Statistical Review 87*(1), 152-177.

Fan, J., S. Upadhye, and A. Worster (2006). Understanding receiver operating characteristic (ROC) curves. *CJEM 8*(1), 19-20.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Garcia-Gutierrez, S., J. M. Quintana, A. Antón-Ladislao, M. S. Gallardo, E. Pulido, I. Rilo, E. Zubillaga, M. Morillas, J. J. Onaindia, N. Murga, R. Palenzuela, and J. G. Ruiz (2017). Creation and validation of the acute heart failure risk score: AHFRS. *Internal and Emergency Medicine 12*, 1197-1206.

Hsu, M.-J. and Y.-H. Chen (2016). Optimal linear combination of biomarkers for multicategory diagnosis. *Statistics in Medicine 35*(2), 202–213.

Iparragirre, A., I. Barrio, and M. X. Rodriguez-Alvarez (2019, 1). On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. *SORT-Statistics and Operations Research Transactions 1*(1), 145–162.

Li, P., J. M. Taylor, D. E. Spratt, R. J. Karnes, and M. J. Schipper (2021). Evaluation of predictive model performance of an existing model in the presence of missing data. *Statistics in Medicine 40*(15), 3477–3498.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association 87*(420), 1227–1237.

Mertens, B. J. A., E. Banzato, and L. C. de Wreede (2020). Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation. *Biometrical Journal 62*(3), 724–741.

Molenberghs, G. and M. G. Kenward (2007). *Missing Data on Clinical Studies*. Wiley.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Quintana, J. M., C. Esteban, A. Unzurrunzaga, S. Garcia-Gutierrez, N. Gonzalez, I. Lafuente, M. Bare, N. F. de Larrea, and S. Vidal (2014). Prognostic severity scores for patients with COPD exacerbations attending emergency departments. *The International Journal of Tuberculosis and Lung Disease 18*, 1415–1420.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. W. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*, 85–95.

Rodrigues, L., P. Bezerra, and L. Saraiva (2008). Morfologia e crescimento dos 6 aos 10 anos de idade em Viana do Castelo, Portugal. *Motricidade 4*.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schetelig, J., L. C. de Wreede, M. van Gelder, N. S. Andersen, C. Moreno, A. Vitek, M. Karas, M. Michallet, M. Machaczka, M. Gramatzki, D. Beelen, J. Finke, J. Delgado, L. Volin, J. Passweg, P. Dreger, A. Henseler, A. van Biezen, M. Bornäuser, S. O. Schönl, N. K. on behalf of the CLL subcommittee, and C. M. W. Party (2017). Risk factors for treatment failure after allogeneic transplantation of patients with CLL: a report from the European Society for Blood and Marrow Transplantation. *Bone Marrow Transplantation 52*, 552-560.

Smith, G. C. S., S. R. Seaman, A. M. Wood, P. Royston, and I. R. White (2014). Correcting for optimistic prediction in small data sets. *Statistical Methods in Medical Research 180*(3).

Steyerberg, E., S. Bleeker, H. Moll, D. Grobbee, and K. Moons (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology 56*, 441–7.

Steyerberg, E. W., F. E. Harrell, G. J. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology 54 8*(8), 774–81.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research 16*(3), 219–242.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*(3), 1–67.

Von Hippel, P. T. (2007). Regression with missing ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology 37*(1), 83–117.

Wahl, S., A.-L. Boulesteix, A. Zierer, B. Thorand, and M. A. van de Wiel (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC Medical Research Methodology 16*.

Wishart, G., C. Bajdik, E. Dicks, E. Provenzano, M. K. Schmidt, M. Sherman, D. C. Greenberg, A. R. Green, K. A. Gelmon, V.-M. Kosma, J. E. Olson, M. W. Beckmann, R. Winqvist, S. S. Cross, G. Severi, D. Huntsman, K. Pylkäs, I. Ellis, T. O. Nielsen, G. Giles, C. Blomqvist, P. A. Fasching, F. J. Couch, E. Rakha, W. D. Foulkes, F. M. Blows, L. R. Bégin, L. J. van't Veer, M. Southey, H. Nevanlinna, A. Mannermaa, A. Cox, M. Cheang, L. Baglietto, C. Caldas, M. Garcia-Closas, and P. D. P. Pharoah (2012). PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. *British Journal of Cancer 107*, 800–807.

Yan, L., L. Tian, and S. Liu (2015). Combining large number of weak biomarkers based on AUC. *Statistics in Medicine 34*, 3811–3830.

Zhu, J. and T. E. Raghunathan (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.